Stats 4CI3 - Assignment 5

Tommy Flynn (400121496)

July 21, 2021

Question 1:

a) The bootstrap estimate of the ratio of the rates of heart attack for the two treatment groups is given by

```
set.seed(13)
asp = c(rep(1,98),rep(0,10939))
pl = c(rep(1,195),rep(0,10839))
M = 1000
t = rep(0,M)
for(i in 1:M){
  temp1 = sample(asp,replace=TRUE)
  temp2 = sample(pl,replace=TRUE)
  t[i] = (sum(temp1)/11037)/(sum(temp2)/11034)
}
mean(t)
```

```
[1] 0.5057724
```

b) A 95% bootstrap percentile interval for the estimate is given by

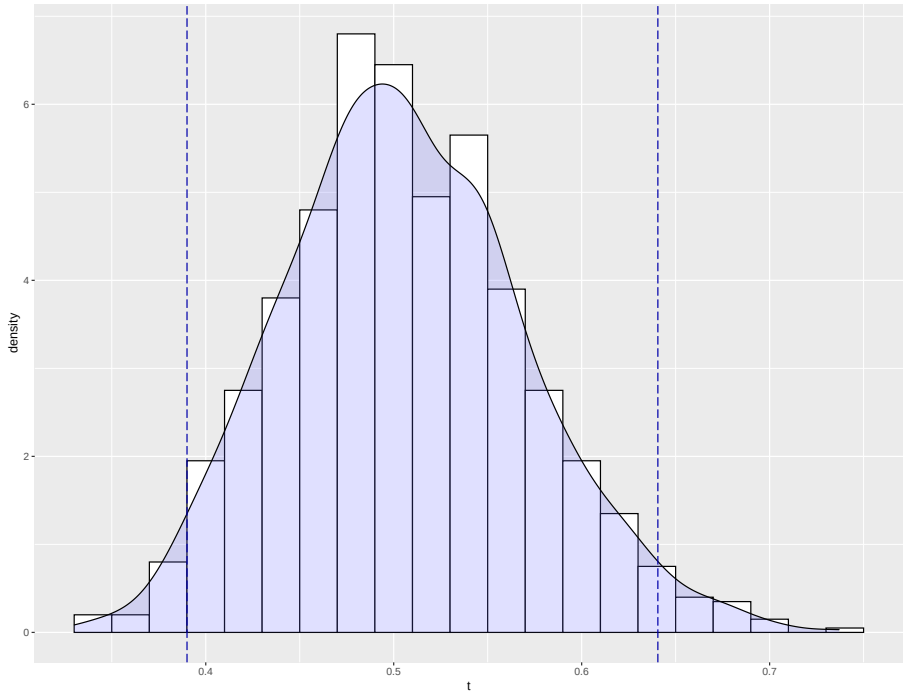
```
sort(t)[0.025*M]
sort(t)[0.975*M]
```

```
[1] 0.3900285
[1] 0.6405444
```

$(0.3900285, 0.6405444)$

c) Here is the histogram with the percentile interval.

```
library(ggplot2)
theta_frame <- data.frame(t)
p <- ggplot(theta_frame, aes(x = t))
p <- p + geom_histogram(aes(y=..density..), binwidth=0.02, colour="black", fill="white
p<-     p + geom_vline(aes(xintercept=sort(t)[0.025*M]), colour="#0000AA", linetype="l
p<-     p + geom_vline(aes(xintercept=sort(t)[0.975*M]), colour="#0000AA", linetype="l
p
```



d) Because we have a an estimate of 0.5 and a 95% percentile interval of $(0.39, 0.64)$, there is evidence that Asprin is preventing heart attacks.

Question 2:

a) The actual coefficient of variation (CV) is

```
x = c(5.32, 9.53, 7.44, 5.71, 6.85, 8.63, 5.98, 6.19, 5.2,
6.81, 8.74, 7.22, 9.22, 6, 6.5, 4.18, 5.12, 7.21, 6.52, 7.31,
12.8, 5.86, 6.82, 6.86, 7.48)
cv = sd(x)/mean(x)
cv
```

```
[1] 0.251914
```

b) The CV using one bootstrap sample is

```
set.seed(13)
b1 = sample(x,replace=TRUE)
sd(b1)/mean(b1)
```

```
[1] 0.1522144
```

c) Here is a 1000 bootstrap sample.

```
M = 1000
b = rep(0,M)
for(i in 1:M){
  temp = sample(x,replace=TRUE)
  b[i] = sd(temp)/mean(temp)
}
```

i. The mean and variance are

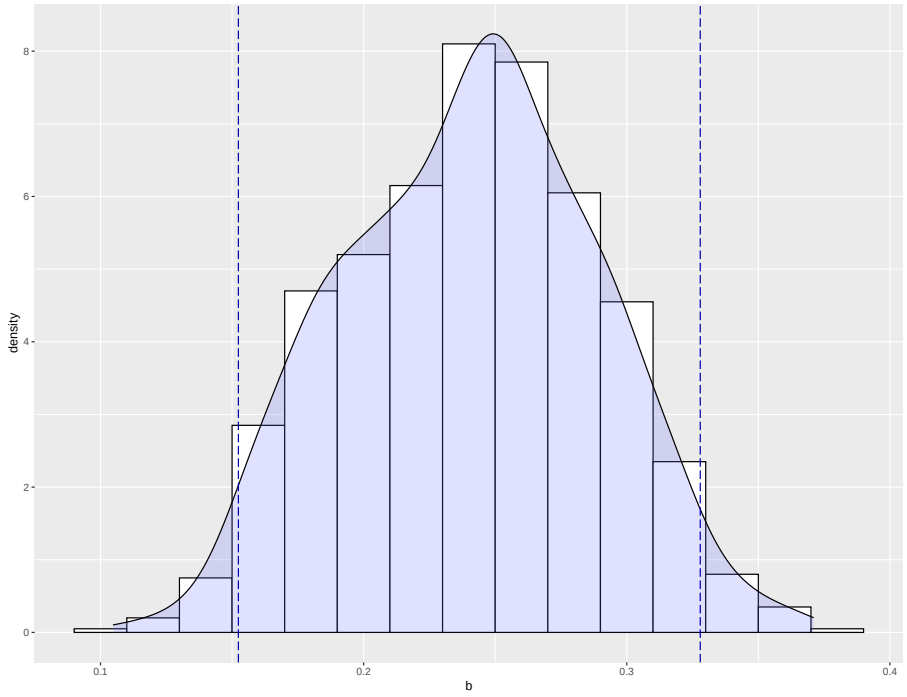
```
mean(b)
var(b)
```

```
[1] 0.2405528
```

```
[1] 0.002279042
```

ii. Here is the histogram.

```
library(ggplot2)
theta_frame2 <- data.frame(b)
p <- ggplot(theta_frame2, aes(x =b))
p <- p + geom_histogram(aes(y=..density..), binwidth=0.02, colour="black", fill="white
p<-     p + geom_vline(aes(xintercept=sort(b)[0.025*M]), colour="#0000AA", linetype="l
p<-     p + geom_vline(aes(xintercept=sort(b)[0.975*M]), colour="#0000AA", linetype="l
p
```



iii. The 95% bootstrap percentile interval is

```
sort(b)[0.025*M]
sort(b)[0.975*M]
```

```
[1] 0.1524042
```

```
[1] 0.3279251
```

$(0.1524042, 0.3279251)$

iv. The bias of the estimate is

```
mean(b) - cv
```

```
[1] -0.01136123
```

Question 3:

a) Here is the Jack-knife estimate.

```
x = c(5.32, 9.53, 7.44, 5.71, 6.85, 8.63, 5.98, 6.19, 5.2,
      6.81, 8.74, 7.22, 9.22, 6, 6.5, 4.18, 5.12, 7.21, 6.52, 7.31,
      12.8, 5.86, 6.82, 6.86, 7.48)
cv = sd(x)/mean(x)

library(bootstrap)
set.seed(13)
theta <- function(x){sd(x)/mean(x)}
results <- jackknife(x,theta)
results

cvj = results$jack.values
mean(cvj)
```

```
[1] 0.2514893
```

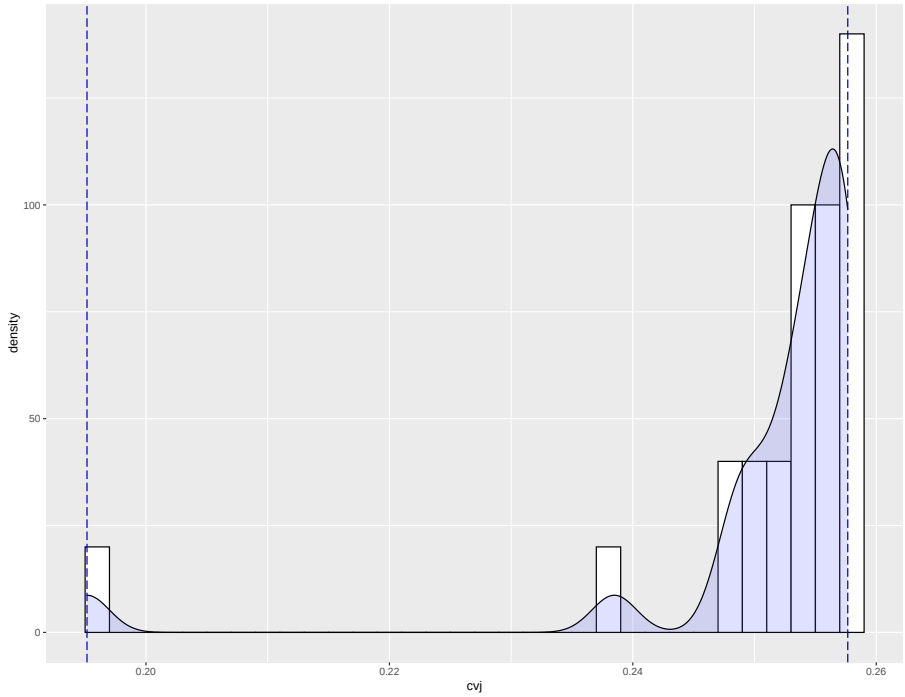
b) The variance of the estimate is

```
var(cvj)
```

```
[1] 0.0001570887
```

c) Here is the histogram.

```
library(ggplot2)
theta_frame2 <- data.frame(cvj)
p <- ggplot(theta_frame2, aes(x =cvj))
p <- p + geom_histogram(aes(y=..density..), binwidth=0.002, colour="black", fill="whit
p<-     p + geom_vline(aes(xintercept=sort(cvj)[1]), colour="#0000AA", linetype="longd
p<-     p + geom_vline(aes(xintercept=sort(cvj)[24]), colour="#0000AA", linetype="long
p
```



d) The 95% percentile interval of the estimate is

```
sort(cvj)[1]
sort(cvj)[24]
```

```
[1] 0.1951539
```

```
[1] 0.2576575
```

$(0.1951539, 0.2576575)$

e) The bias of the estimate is

```
mean(cvj) - cv
```

```
[1] -0.0004247377
```

Question 4:

a) The posterior distribution of parameter θ is created using the data and the prior distribution. Specifically, it is constructed using Bayes formula stating the posterior distribution is equal to the likelihood times the prior over the evidence. That is,

$$\pi(\theta|x_1,\ldots,x_n) = \frac{L(\theta|x_1,\ldots,x_n)\pi(\theta)}{\int \pi(\theta)L(\theta|x_1,\ldots,x_n)d\theta}$$

Where
$\pi(\theta|x_1,\ldots,x_n)$ is the posterior distribution
$\pi(\theta)$ is the prior distribution
$L(\theta|x_1,\ldots,x_n)$ is the likelihood
$\int \pi(\theta)L(\theta|x_1,\ldots,x_n)d\theta$ is the evidence

The evidence is a constant that ensures that the posterior distrbution is a probability density function so it is often written as

$$\pi(\theta|x_1,\ldots,x_n) \propto L(\theta|x_1,\ldots,x_n)\pi(\theta)$$

b) The Monte Carlo Bayesian Inference Technique is where one estimates using a sample simulated from the derived posterior distribution.

c) A Monte Carlo estimater of $\mathrm{Var}(\theta|x_1,\ldots,x_n)$ for the posterior distribution is

$$\widehat{\mathrm{Var}}(\theta|x_1,\ldots,x_n) = \frac{1}{N}\sum_{i=1}^{N}\theta_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\theta_i\right)^2$$

d) A $100(1-\alpha)\%$ Bayesian Credible Interval is an interval estimater of parameter θ in the sense that the likelihood of θ belonging to that interval is $100(1-\alpha)\%$. One can compute a credible interval by taking the lower and upper quantiles of the posterior distribution which depends on the prior and α.

e) A Non-informative prior is a prior distribution that is designed to minimize the effect on the final inference.

Question 5:

a) The ggs-density function takes in markov chains and outputs the density plots of those chains by colour.

b) The ggs-compare-partial function takes in markov chains and provides density plots comparing the distributions of the chains with only its last part.

c) The ggs-traceplot function takes in markov chains and provides a time series plot of the chains.

d) The ggs-running function takes in markov chains and provides a plot of the running means.

e) The ggs-autocorrelation function takes in markov chains and provides a plot of the autocorrelation matrix between the chains.

d) The ggs-caterpillar function takes in markov chains and provides a caterpillar plot by combining all the chains for each parameter.